Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 61 (2015) 367 – 372

Procedia
Computer Science

Complex Adaptive Systems, Publication 5
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2015-San Jose, CA

Synthetic Rating on Talent Evaluation-Similarity of Subsets

Shijun Tang[a], Rajan Alex[b*]

[ab]*Department of Engineering and Computer Science School of Engineering, Computer Science, and Mathematics*
West Texas A&M University, Canyon, TX 79016, USA

Abstract

There are many topics about rating individuals, animals, places, things, or abstract ideas that are actively researched. When rating about an object is needed to form an opinion it is often given by an expert in the field. These ratings vary from one individual expert's rating to another due to the subjective nature of the evaluation process. How can we evaluate the ratings? How can we find the correlations and similarities among these sets? How can we provide a mathematical modelling for a rating problem? This paper provides a procedure for the extension of fuzzy synthetic rating modelling on a sample to the entire data set and introduces a *k*-means clustering method to check the level to which there exist similarities among the subsets and classify the dataset automatically for a rating problem. The related synthetic rating and an example to illustrate the modelling is given in this work.

Peer-review under responsibility of scientific committee of Missouri University of Science and Technology

Keywords: synthetic rating; k-means clustering; fuzzy sets.

1. Introduction

Human beings can easily perform synthetic evaluation from several different features or qualities and make meaningful decisions intuitively. However, these decisions are generally subjective in nature; there is no known precise formula that our brain can apply to arrive at conclusions. The conclusions often vary from one evaluator to another. There is always the possibility of 'human bias' in such decision making processes although we are in most situations willing to accept the decisions. For example, the decision to hire a faculty member for a university teaching/research position is not always based on merits and qualifications but other factors such as how well is the prospective candidate a good 'fit' for the position in the department? There is no one formula the search committee can apply to make this decision although such decisions are made routinely. The decision making problems where

*Correspodning Author, Tel.: 1-806-651-2288; fax: 1-806-651-5259
E-mail address: ralex@wtamu.edu

Peer-review under responsibility of scientific committee of Missouri University of Science and Technology
doi:10.1016/j.procs.2015.09.162

human subjectivity is involved do not have a formula or model. It is fuzzy! In general, the problem of synthetic rating does not have a mathematical model representation.

Suppose a company wants to evaluate its employees for merit raises, the decision always involves some amount of subjectivity. How can we evaluate an employee's qualities as good for a merit raise? The good quality an employee must possess is complex and cannot be evaluated by one single value. However, we can decompose good qualities into several factors such as knowledge, skill, experience, dependability, and so on. Each of these factors can then be assigned a quantifying numeric value and synthetic rating performed. A mathematical formulation methodology for synthetic rating may be found in [3], [5], and [6].

The synthetic rating work given in [6] uses a fuzzy synthetic rating process involving fuzzy clustering analysis using theoretical framework on fuzzy regression and fuzzy neural networks [1], [2] and [4]. The fuzzy synthetic rating problem is a mapping from a given factor space *F* into a fuzzy score space *S*. We propose a mapping from a subset of the factor space consisting of sample points and extending the mapping on the subset to the entire factor space *F*. To build the mapping, we will employ clustering of the sample points using the *K*-mean process. This work will explore the synthetic rating problem. In section 2, we will give an example and set up the synthetic rating problem. In section 3, we will give the clustering process and in section 4, we will give the mapping extension process and apply it to an example problem. In section 5, we will give the conclusion and further research directions for this work.

2. Sampling and subjective ratings by experts

The synthetic rating problem is explained in detail in this section using an example. Assume that there is a need to decide about the promotions and salary raises for the employees in a company. To evaluate each employee during the evaluation month for the employee, the company has a rating system that gives a rating record of each employee of the company according to the following six evaluation criteria or sub-indices: skill, knowledge, hardworking, responsible, always on time, and courtesy. For every k^{th} employee of the company, $k = 1,2,3,\cdots$, there are the rating records as follows:

x_{k1} - skill,
x_{k2} - knowledge ,
x_{k3} - hardworking,
x_{k4} - responsible,
x_{k5} - always on time, and
x_{k6} - courtesy.

Using the above rating criterions, how can we get the synthetic evaluation for each employee of the company centered on his/her good qualities and talents? First we need to get experts who are familiar with the evaluation process to give their numeric representation for each of the indices for a chosen sample of employees. Assume that we have a sample size of 12 employees to be evaluated by an expert with respect to the mentioned 6 evaluation indexes. Here we denote the synthetic evaluation index as 'good employee qualities.' We note that the number of sub-indices and what each

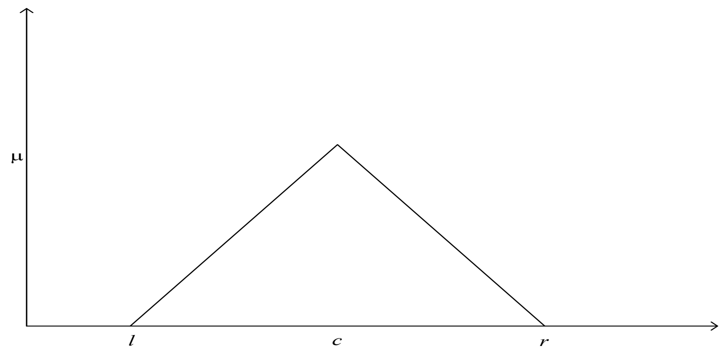

Figure 1 Triangular membership function

one represents will vary for each situation and that the above list of sub-indices is only a sample list of sub-indices for this evaluation methodology. Suppose further that the expert's evaluation of each sub-index given as a subset of R^6 is as follows:

$$x_1 = (0.8, 0.7, 0.9, 0.8, 0.8, 0.7),$$
$$x_2 = (0.9, 0.7, 0.9, 0.9, 0.8, 0.9),$$
$$x_3 = (0.9, 0.8, 0.8, 0.8, 0.7, 0.8),$$
$$x_4 = (0.5, 0.3, 0.4, 0.5, 0.8, 0.8),$$
$$x_5 = (0.3, 0.4, 0.5, 0.5, 0.7, 0.9),$$
$$x_6 = (0.4, 0.5, 0.3, 0.4, 0.9, 0.7),$$
$$x_7 = (0.9, 0.7, 0.8, 0.7, 0.4, 0.3),$$
$$x_8 = (0.7, 0.9, 0.8, 0.7, 0.3, 0.2),$$
$$x_9 = (0.3, 0.4, 0.5, 0.4, 0.4, 0.3),$$
$$x_{10} = (0.5, 0.3, 0.4, 0.4, 0.3, 0.4),$$
$$x_{11} = (0.5, 0.6, 0.7, 0.6, 0.5, 0.6),$$
$$x_{12} = (0.6, 0.5, 0.7, 0.7, 0.6, 0.5).$$

Here each sub-index is represented by a number in the unit interval $[0,1]$. It forms a group of sampling points. Each scored rating number could also come from a set such as $\{1,2,3,4,5\}$, $\{1,2,\ldots,10\}$, or $\{1,2,\ldots,100\}$. In such situations, we can transfer any set of discrete values into values in the unit interval $[0,1]$, where 1 stands for the highest discrete value (excellent rating) and 0 stands for the lowest discrete value (poor rating) that could be assigned. The group of employees is evaluated by the same expert with respect to a synthetic evaluation index. Since the synthetic evaluation is fuzzy, in this work, we will assume that the expert's rating for each employee is a triangle fuzzy number. Assuming that the evaluator giving the rating is not an expert in fuzzy logic, we can get a triangular fuzzy number by asking the following questions: what is the number the evaluator would be most comfortable to assign as his/her rating for 'good employee qualities'? The evaluator is asked to give a central value *c*; then we can ask: what is the lower limit of your rating? The evaluator gives the value *l*; and we can ask: what is the upper limit of your rating? The evaluator gives the value *r*. Thus we get the triangle fuzzy number (*l, c, r*) as given in Figure 1. More details about this representation may be found in [1] and [2]. Assume that the 12 employees' synthetic ratings by the expert are records as follows:

$$y_1 = (0.78, 0.8, 0.85),\ y_2 = (0.87, 0.90, 0.91),\ y_3 = (0.82, 0.85, 0.89),$$
$$y_4 = (0.65, 0.70, 0.80),\ y_5 = (0.60, 0.65, 0.75),\ y_6 = (0.50, 0.6, 0.63),$$
$$y_7 = (0.50, 0.55, 0.57),\ y_8 = (0.56, 0.59, 0.63),\ y_9 = (0.3, 0.31, 0.33),$$
$$y_{10} = (0.35, 0.40, 0.44),\ y_{11} = (0.50, 0.58, 0.63),\ y_{12} = (0.52, 0.53, 0.56).$$

To perform synthetic rating, we first have to find some similarities or patterns to categorize the representation; we employ clustering on the sample.

3. Clustering based on K-Means algorithm

A well-known clustering *algorithm,* k-means is one of the simplest unsupervised learning algorithms. In the paper, k-means algorithm is used to find clusters and get the evaluation of rating. The main idea is to define *k* centroids, one for each cluster. These centroids should be put as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. We need to re-calculate *k* new centroids as centroid of the clusters resulting from the previous step.

Once we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid iteratively. As a result, we may notice that the k centroids change their location step by step until no more changes are done. Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function is

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - c_j\right\|^2,$$

where $\left\|x_i^{(j)} - c_j\right\|^2$ is a chosen Euclidian distance measure between a data point $x_i^{(j)}$ and the cluster center c_j. This is an indicator of the distance of the n data points from their respective cluster centers.

The algorithm may be summarized in the follows steps:

1. Design k points which are initial group centroids.
2. *Classify each point to the group that has the closest centroid.*
3. *Calculate the positions of the k centroids again.*
4. *Repeat Steps 2 and 3 until the centroids no longer move.*

We can construct a matrix A by putting all sub-indices given as a subset of R^6 together when searching for clusters or similarities among set components (subsets) in the whole set.

$$A = \begin{bmatrix} A_1 \\ \vdots \\ Am \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix},$$

where A is a matrix including m rating sets. For the example, we put the sample size of 12 employees into the matrix $\boldsymbol{A}$ given by

$$A = \begin{bmatrix} 0.8 & 0.7 & 0.9 & 0.8 & 0.8 & 0.7 \\ 0.9 & 0.7 & 0.9 & 0.9 & 0.8 & 0.9 \\ 0.9 & 0.8 & 0.8 & 0.8 & 0.7 & 0.8 \\ 0.5 & 0.3 & 0.4 & 0.5 & 0.8 & 0.8 \\ 0.3 & 0.4 & 0.5 & 0.5 & 0.7 & 0.9 \\ 0.4 & 0.5 & 0.3 & 0.4 & 0.9 & 0.7 \\ 0.9 & 0.7 & 0.8 & 0.7 & 0.3 & 0.2 \\ 0.3 & 0.4 & 0.5 & 0.4 & 0.4 & 0.3 \\ 0.5 & 0.3 & 0.4 & 0.4 & 0.3 & 0.4 \\ 0.5 & 0.6 & 0.7 & 0.6 & 0.5 & 0.6 \\ 0.5 & 0.6 & 0.7 & 0.6 & 0.5 & 0.6 \\ 0.6 & 0.5 & 0.7 & 0.7 & 0.6 & 0.5 \end{bmatrix}$$

In the paper, we have chosen the four and five clusters (k = 4, 5), and used the Matlab code. The advantage of this approach for clustering analysis is that the user can decide on how many clusters for the sample data he wants to have.

For instance, after running the k-means codes, we have obtained the fuzzy clusters as the following:

$$C_1 = \{x_1, x_2, x_3\},\ C_2 = \{x_4, x_5, x_6\},\ C_3 = \{x_7, x_8, x_{11}, x_{12}\},\ C_4 = \{x_9, x_{10}\} \text{ when } k = 4 \text{ and}$$

$$C_1 = \{x_1, x_2, x_3\},\ C_2 = \{x_4, x_5, x_6\},\ C3 = \{x_7, x_8\},\ C_4 = \{x_9, x_{10}\},\ C_5 = \{x_{11}, x_{12}\} \text{ when } k = 5$$

It is worth noting that when k = 5, the clustering classes agree with an earlier work [6] by a different approach. However, the flexibility of this method is that the user gets to choose the value of k for the number of clusters needed for the modeling.

For the remaining work we will use the k = 4 clustering classes. Now we propose to find the center for each clustering class by taking the average of the ratings for each of the classes. Thus the center of each class is given as follows:

$$c_1^* = (0.87, 0.73, 0.87, 0.83, 0.77, 0.80),$$
$$c_2^* = (0.40, 0.40, 0.40, 0.47, 0.80, 0.80),$$
$$c_3^* = (0.68, 0.68, 0.75, 0.68, 0.45, 0.40),$$
$$c_4^* = (0.40, 0.35, 0.45, 0.40, 0.35, 0.35).$$

Similarly, the synthetic rating of each class is also gotten by arithmetic averaging as:

$$y_1^* = (0.82, 0.85, 0.88)$$
$$y_2^* = (0.58, 0.65, 0.73)$$
$$y_3^* = (0.52, 0.56, 0.60)$$
$$y_4^* = (0.33, 0.36, 0.39).$$

Since the central point of each class reflects the evaluation index of a typical object, the synthetic rating of each class is called the typical synthetic rating. We can use the synthetic rating classes to rate an object of the space from which the sample was initially taken as follows:

4. Extension of the rating on the sample to the space

Suppose we have m cluster classes, with center ratings $c_1^*, c_2^*, \ldots, c_m^*$ and synthetic ratings $y_1^*, y_2^*, \ldots, y_m^*$, we define distances for an arbitrary employee rating $x_0 = (x_{01}, x_{02}, \ldots, x_{0n})$ as

$$d_j = \sum_{k=1}^{m} | c_{jk} - x_{0k} |,$$

where $c_j^* = (c_{j1}, c_{j2}, \ldots, c_{jn})$, n = the number of sub-indices. Then define

$$d^* = \max_j \{ \lceil d_j \rceil, j = 1,2,\ldots,m \}, \text{ where } \lceil \cdot \rceil \text{ denote the ceil function.}$$

$$h_j^* = \frac{d^* - d_j}{d^*}, \quad j = 1,2,\ldots,m, \text{ and}$$

$$f_j^* = \frac{h_j^*}{\sum_{j=1}^{m} h_j^*}, \quad j = 1,2,\ldots,m$$

Using the f_j^* we can get an extension mapping with center, left boundary, and right boundary as follows.

$$c^* = \sum_{j=1}^{m} f_j^* c_j^*, \quad l^* = \sum_{j=1}^{m} f_j^* l_j^* \text{ and } r^* = \sum_{j=1}^{m} f_j^* r_j^*, \text{ where } y_j^* = (l_j^*, c_j^*, r_j^*), \quad j = 1,2,\ldots,m$$

Assume that the expert's rating for the sub-indices is $x^* = (0.5, 0.4, 0.5, 0.5, 0.6, 0.5)$. For the above definitions, and the clustering sample as in section 4, we get

$$d_1 = 1.87, \; d_2 = 0.73, \; d_3 = 1.14, \; d_4 = 0.7 \text{ and}$$
$$d^* = 2,$$

and the triangle membership function with left boundary, center, and right boundary as

$$y_0 = (0.48, 0.53, 0.58).$$

Next we further analyze the arbitrarily chosen data point. For this, let us now take $x^* = (0.5, 0.4, 0.5, 0.5, 0.6, 0.5)$ as x_{13}, the 13th sample point and apply the test set to the matrix A and use the k-means algorithm again. When setting $k = 4$, for four clusters, we obtain new clusters as $C_1 = \{x_1, x_2, x_3\}$, $C_2 = \{x_4, x_5, x_6\}$, $C_3 = \{x_7, x_8, x_{11}, x_{12}\}$, $C_4 = \{x_9, x_{10}, x_{13}\}$. We note that the test set x_{13} belongs to the fourth cluster with the sample points x_9, x_{10}. In the cluster C_4, the center data is $c^* = (0.4333, 0.3667, 0.4667, 0.4333, 0.4333, 0.4000)$.

We can observe the similarity relationship from figure 2 as follows: the indices x_{k1} - skill, x_{k2} - knowledge, x_{k3} - hardworking, are exactly overlapping with that of x_9 or x_{10} sample point, only x_{k4} - responsible, x_{k5} - always on time, and x_{k6} - courtesy, are slightly higher than those in x_9 or x_{10} sample point. But the deviations for the criteria x_{k4}, x_{k5}, and x_{k6} from its center are symmetric and reasonable. Therefore, the Figure 2 confirms that we obtained the same results using two different methods, k-means algorithm and, our synthetic rating method [3]. We also infer that the test set falls in the same category as obtained by $y_0 = (0.48, 0.53, 0.58)$. Furthermore, Figure 2, suggests that the test set has good similarity with the set x_9 and x_{10}.

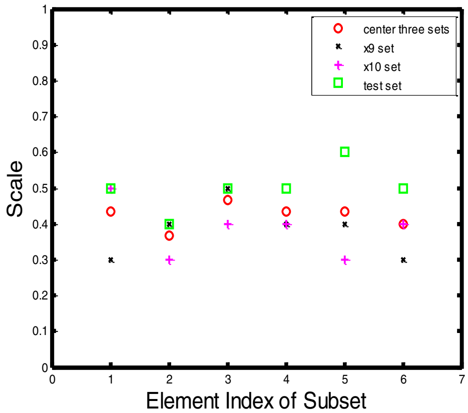

Figure 2 Similarity relationships among the subsets in cluster C_4

5. Conclusion

A simple synthetic rating evaluation process is suggested here. In this work, we have provided an extension of the fuzzy synthetic rating on a sample to the entire data set under consideration. We have illustrated the fuzzy synthetic rating process through an application problem scenario. We have employed the *k*-mean clustering process to get the clustering among the sample points. Also, we verify that the *k*-means cluster process produced a synthetic rating result that agrees with our work. As future work, the authors would like to explore novel clustering techniques, synthetic rating, and provide theoretical formulation to support the comparison of an arbitrary data point of the space under consideration with the sample cluster that contains the data point.

References

1. Alex, R., 2004, “Fuzzy normal regression model and related neural networks,” Soft Computing, Vol. 8 (10), pp. 717 – 721
2. Alex, R., 2006, “A new kind of fuzzy regression modelling and its combination with fuzzy inference,” Soft Computing, Vol. 10 (7), pp. 618-621
3 .Alex, R., 2008, “Synthetic rating on talent evaluation,” Proceedings of the ANNIE Conference 2008 Vol. 18, 2008, 757-762.
4. Cheng, C.B., Lee, E.S., 2001, “Fuzzy regression with radial basis function network,” Fuzzy Sets and Systems, Vol. 119, pp. 291-301.
5. Granath, G., 1984, “Application of fuzzy clustering and fuzzy classification to evaluate the provenance of glacial till,” Mathematical Geology, Vol. 16, pp. 283-300
6. Lee, H., Tanaka, H., 1999, “Fuzzy approximations with non-symmetric fuzzy parameters in fuzzy regression analysis,” Journal of the Operations Research Society of Japan, Vol. 42, pp. 98-112